DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES

CALIFORNIA INSTITUTE OF TECHNOLOGY

PASADENA, CALIFORNIA 91125

Uncovering Behavioral Strategies:
Likelihood-Based Experimental Data Mining

Mahmoud A. El-Gamal and David M. Grether

SOCIAL SCIENCE WORKING PAPER 850

July 1993

Uncovering Behavioral Strategies: Likelihood-Based Experimental Data-Mining

Mahmoud A. El-Gamal and David M. Grether *

DHSS 228-77, Caltech, Pasadena, CA 91125.

Revised: July 23, 1993

Abstract

Economists and psychologists have recently been developing new theories of decision making under uncertainty that can accommodate the observed violations of standard statistical decision theoretic axioms by experimental subjects. We propose a procedure which finds a collection of decision rules that best explain the behavior of experimental subjects. The procedure is a combination of maximum likelihood estimation of the rules together with an implicit classification of subjects to the various rules, and a penalty for having too many rules. We prove that our procedure yields consistent estimates (as the number of tasks per subject, and the number of subjects go to infinity) of the number of rules being used, the rules themselves, and the proportion of our population using each of the rules. We apply our procedure to data on probabilistic updating by subjects in four different universities. We get remarkably robust results which show that the most important rules used by the subjects are Bayes rule, representativeness rule (ignoring the prior), and conservatism (over-weighting the prior). In our procedure, the subjects are allowed to make errors, and our estimated error rate is typically 20%.

*The first coauthor acknowledges financial support from NSF grant #SES-9223701 to the California Institute of Technology. We thank the Jet Propulsion Laboratory for giving us access to their Cray YMP, on which all of our calculations were made.

Uncovering Behavioral Strategies: Likelihood-Based Experimental Data-Mining

Mahmoud A. El-Gamal and David M. Grether

1 Introduction

The economic theory of decision making under uncertainty has been seriously challenged by a series of discoveries of violations of that theory by experimental subjects. The paradoxes of Allais (1953) and Ellsberg (1961) are among the earliest examples, but recently, psychologists have added several others. Among the more recent violations of statistical decision theory are framing effects, certainty effects, common ratio effects (Kahneman and Tversky (1979)), preference reversals (Lichtenstein and Slovic (1971), Grether and Plott (1979)), to name but a few. Part of the economists' response to these developments has been the introduction of a number of new theoretical models of decision making designed to be consistent with some of the reported violations of expected utility theory. Among those theories are Machina (1982)'s "fanning out". (Loomes and Sugden (1987), Bell (1982))'s regret theory, Chew (1983)'s generalized weighted utility theory, (Quigen (1982), Yaari (1987))'s rank dependent utility, and Kahneman and Tversky (1979)'s prospect theory.

While economists have been introducing new models of individual decision making, psychologists have developed a number of heuristic explanations of specific individual behaviors. For preference reversals, those models include Lichtenstein and Slovic (1971)'s anchoring and adjustment, Goldstein and Einhorn (1987)'s expression theory, Bostic et al. (1990)'s response mode, Loomes et al. (1989)'s regret, Tversky et al. (1988)'s contingent weighting theory, Mellers et al. (1992)'s change of process theory, Birnbaum et al. (1992)'s configural weighting theory. Research on judgement of probabilities has produced an array of heuristics which individuals use in different circumstances (c.f. Tversky and Kahneman (1974)). Prominent heuristics include Kahneman and Tversky (1972)'s representativeness, Tversky and Kahneman (1972)'s availability, Edwards (1982)'s conservatism, and Lichtenstein and Slovic (1971)'s anchoring and adjustment. Recent research (e.g. Payne (1982), Gigerenzer et al. (1991)) suggests that the dependence of such heuristics on the specific context of the decision making experiment is not fully understood.

One clear suggestion of the above referenced literature is the need to develop theories of decision making under uncertainty which are built on an empirically justified foundation. The literature in its current state does not support the conclusion that economic subjects are sufficiently homogeneous to be described by a single theory. Different subjects may use different decision rules, and if the rules they use do not yield satisfactory outcomes, they may abandon them and use different ones (c.f. Mellers et al. (1992)). In this paper, we devise and use a general estimation/classification procedure which uncovers the most likely collection of rules that experimental subjects use. The procedure is quite simple in nature. We start by choosing a class of decision rules that the subjects may use. The class of decision rules provides us with a class of likelihood functions for each data point. For a fixed number of rules in our class, we estimate the maximum likelihood collection of rules that the agents are using, as well as the proportions of our population using each of the rules. By adding an information criterion that penalizes our procedure for admitting more rules, our procedure achieves consistent estimates (as the number of tasks that each individual performs in the experiment, and the number of individuals, go to infinity) of the number of rules being used, the rules themselves, and the proportion of the population using each of the rules.

It is clear that our general procedure is applicable in a variety of situations where a class of data generating processes is considered, and our sample is not assumed to be generated by only one of those data generating processes. The applications of our procedure are not restricted to decision making experiments, or even to experimental settings. However, our consistency results use the fact that in experimental settings, we can let the number of tasks per individual, and the number of individuals, get large to obtain estimates within a required tolerance level. In environments where a finite data sample is given, we typically cannot choose the design and the sample sizes to get our estimates of the collection of data generating processes within a pre-specified tolerance of their true values. We emphasize the fact that finite sample bias and misclassification corrections may be needed when analyzing given experimental data, but that issues of optimal experimental design should be seriously considered to minimize the need for such corrections.

We close this introduction by briefly comparing our approach to the vast and growing literature on classification and clustering. The main difference that we want to point out is that a primary goal of the classical classification literature is to establish simple algorithms that work for a very large class of problems. For example, Wallace and Boulton (1968)'s Snob (and later Snob 2) program, and Cheeseman (1988)'s Autoclass II program assume normality of the data generating process. In contrast, our approach does not rely on a custom-made algorithm. The general procedure we use agrees with all likelihood-based classification procedures in its form, but the class of likelihood functions is suggested by the problem. In that sense, we are closer to the modern coding theoretic approach to estimation and classification (e.g. Wallace and Freeman (1987), Rissanen (1987)). In that literature, the class of likelihood functions is used as a means for summarizing the important aspects in a data set. The trade-off between using simpler classes of likelihood functions and explaining the data better is explicitly treated in that literature. In our framework, where the number of tasks per experimental subject, and the number of subjects, go to infinity, we prove that any strictly increasing penalty function (with bounded increments) for the complexity of our model will yield consistency of our estimates of the number of classes, the classes themselves, and the proportions of the population in each of the classes. Specific penalty functions from the literature (e.g. Akaike (1974)'s famous criterion) will work. We derive a particular pseudo-Bayesian penalty function that seems appropriate for our example.

The remainder of the paper will proceed as follows. In section 2, we describe the collection of experiments that we analyze in this paper. In section 3, we discuss the difficulties encountered in a brute-force classification approach, and introduce a class of decision rules that reduces the computational burden to feasible levels. In section 4, we present our likelihood-based estimation/classification procedure for the particular application at hand, and motivate a particular penalty function for the number of classes allowed. In section 5, we state our estimation/classification procedure in its general form which is applicable in many more contexts, and prove the consistency of our estimates of the number of classes, the classes themselves, and the proportion of the population belonging to each of the classes. In section 6, we discuss the methods used for implementing our procedure in our application, and discuss the results that we obtain from the experiments described in section 2. In that section, we also address some of the suboptimal properties that our procedure may possess in finite samples, and discuss how they can be ameliorated by considerations of optimal experimental design. Section 7 concludes the paper.

2 The Experiments

The experimental data that we use in this paper were collected at four different educational institutions. Subjects were recruited from economics classes at UCLA, Occidental College, California State University at Los Angeles, and Pasadena City College. The subjects were told that they were to participate in an economics experiment, and that they would be payed for their participation. Upon arrival, the subjects were randomly divided into two groups. The procedures were identical for both groups except for the method of payment. The two groups performed the experimental tasks independently in two different rooms.

In each room, there were three bingo cages. One cage contained **six** balls numbered one through six. this cage was used to generate the priors which were of the form: **if one through m is drawn, use cage A, otherwise use cage B**. Cages A and B were the populations from which the observations were drawn:

Cage A contained six balls, four labelled N and two labelled G.

Cage B also contained six balls, three with each letter.

At the beginning of each experiment, the instructions were read and the subjects elected one person to serve as a monitor. The monitors inspected all equipment, observed the draws from the cages and, generally, checked to be sure that the experimenters were being truthful. The monitors did not communicate with the subjects outside of their duties as monitor. The monitors were guaranteed a payment at least equal to the average received by subjects in their rooms.

The experiment proceeded as follows. Cages A and B were placed behind an opaque screen, and a prior was announced. The prior cage was spun and a ball selected thus determining whether cage A or cage B would be used. The three prior rules used were two, three or four chances out of six for cage A. The cage selected was placed in the front of the room and six draws (with replacement) were performed, and the results were announced and written on a blackboard. Subjects also recorded the outcomes on their answer sheets. Subjects were then asked to indicate which cage they believe was used in generating the observations. After all subjects had indicated which cage they felt was the more likely, a new prior was announced, and the procedures were repeated.

In one room, all subjects were payed a flat fee. In the other room, one task was selected (randomly, using a bingo cage), and subjects earned a \$10 bonus if their response was correct. A response was considered correct if the cage the subject stated was the more likely was in fact the cage from which the balls were drawn. The sessions lasted approximately one and one half hours, and the number of decisions made by each subject ranged from 14 to 21. The raw data is reported in Tables 1-8 of Appendix B. The instructions and a sample answer sheet for experimental subjects are shown in Appendix C.

3 A Natural Class of Decision Rules

Ignoring the order of the draws, there are seven possible outcomes of the draws (zero through six N's) and three priors, resulting in 21 possible decision situations. In each of these situations, the subjects could chose either cage A or cage B. Therefore, there are in principle $2^{21} = 2,097,152$ possible decision rules. This is a very large number, and, fortunately, most of those rules can be excluded by very simple rationality considerations. Since cage A has a higher proportion of N's than cage B, the outcome most strongly favoring cage B would be no N's (six G's), and the one most strongly favoring cage A would be six N's (no G's). A natural rule would be to have a cutoff number for each of the priors, such that if the number of N's exceeds that cutoff number, the rule selects cage A, otherwise it selects cage B.

By imposing the cutoff class of rules discussed above, we need to decide how to treat the behavior of a subject who is observed choosing cage B when some number of N's has been observed, and then choosing A when a smaller number has been observed. We shall introduce the possibility of experimental subjects making errors (i.e. deviating from the rule). This will allow each of our decision rules (which for our data will simply be a triple of numbers between zero and six) to give a positive probability (likelihood) to all possible patterns of behavior. We shall assume that each subject uses a decision rule (c_1, c_2, c_3), of the form: under prior i, choose cage A if the number of N's observed is greater than c_i, and choose cage B otherwise. With probability ϵ, however, the subject trembles and makes a random choice. In other words, for each decision (given a prior, and a number of N's drawn) with probability $(1-\epsilon)$ the subject follows the rule (c_1, c_2, c_3), and with probability ϵ the subject chooses cage A with probability 1/2 and cage B with probability 1/2. Now, the number of possible rules $\{(c_1, c_2, c_3); -1 \leq c_i \leq 6; i = 1,2,3\}$ is $8^3 = 512$ (where c_i's are integers, and we use -1 as the lower bound corresponding to always choosing cage A, even if zero N's were observed). If we further restrict the set of rules by requiring that $c_i \geq c_j$ if $i \geq j$ (i.e. given a stronger prior in favor of cage B, a subject should require at least as many N's to choose cage A), there are only 84 possible rules.

There is a potential identification problem which arises because of the possibility of agents making errors. We could model randomness by having individuals use a rule with probability $(1-x)$, and deviate with probability x. In that framework, consider a subject who for the same prior and outcome chooses cage A $100x$ percent of the time and B $100(1-x)$ percent of the time. The subject could be seen as choosing cage A but making a mistake $100(1-x)$ percent of the time, or choosing cage B but making a mistake $100x$ percent of the time. Thus, for each of our 2^{21} conceivable decision rules, together with an error rate, there is another rule (the opposite of the first), and another error rate (one less the first) which is observationally equivalent to that rule. Notice that by imposing the restriction to the class of cutoff rules described above, this identification problem does not arise since those opposite decision rules are rendered inadmissible (except for the rules that always pick cage A, or always pick cage B).

4 A Likelihood-Based Estimation/Classification Procedure

As stated in the previous section, we have restricted attention to a class of decision rules which can be written as (c_1, c_2, c_3) where c_i is the cutoff rule used when prior i is induced. Prior 1 corresponds to cage A having probability 1/3, prior 2 to cage A having probability 1/2, and prior 3 corresponds to cage A having prior probability 2/3. We assume that each of our subjects uses one such rule (c_1^s, c_2^s, c_3^s) from the class $\mathfrak{C} = \{(c_1, c_2, c_3) : -1 \leq c_i \leq 6; i = 1, 2, 3\}$. As we have already pointed out, the class $\mathfrak{C}$ has $8^3 = 512$ elements. We shall allow different subjects to be using different rules. We further assume that the error rate ϵ discussed above is the same for all subjects, and all tasks. For each rule $c = (c_1, c_2, c_3)$, and error rate ϵ, we get a likelihood function $f^{c,\epsilon} \in \mathfrak{F}$, where the class $\mathfrak{F}$ consists of all likelihood functions for $(c, \epsilon) \in \mathfrak{C} \times [0, 1]$. For a subject s, given a sequence of observations $x_1^s, \ldots, x_{t_s}^s$, where $x_\tau^s = (p_\tau, N_\tau, a_\tau)$, $p_\tau \in \{1, 2, 3\}$ is the prior, N_τ is the number of N's observed, and a_τ is the choice (A or B) of the subject, define the variable

$$x_{c,\tau}^s = \begin{cases} 1 & \text{if } (a_\tau{=}\text{A and } N_\tau > c_{p_\tau}) \text{ or } (a_\tau{=}\text{B and } N_\tau \leq c_{p_\tau}); \\ 0 & \text{otherwise} \end{cases}$$

Now define the sufficient statistic $X_c^s = \sum_{\tau=1}^{t_s} x_{c,\tau}^s$. Then under rule $c = (c_1, c_2, c_3)$, and error rate ϵ, the likelihood function $f^{c,\epsilon}$ can easily be computed as follows:

$$f^{c,\epsilon}(x_1^s, \ldots, x_{t_s}^s) = \left(1 - \frac{\epsilon}{2}\right)^{X_c^s} \times \left(\frac{\epsilon}{2}\right)^{t_s - X_c^s}.$$

Now, we observe data on n experimental subjects, with each subject s being observed over t_s tasks. If we assume that all agents are using the same rule $c \in \mathfrak{C}$, then we can estimate (c, ϵ) by the maximum likelihood estimates:

$$(\hat{c}, \hat{\epsilon}) = \text{argmax}_{c,\epsilon} \prod_{s=1}^{n} f^{c,\epsilon}(x_1^s, \ldots, x_{t_s}^s).$$

If we assume that different agents may be using different rules, and that there are exactly k such rules $c^1 = (c_1^1, c_2^1, c_3^1), \ldots, c^k = (c_1^k, c_2^k, c_3^k)$, then we can estimate $c^1, \ldots, c^k, \epsilon$ by

$$(\hat{c}^1, \ldots, \hat{c}^k, \hat{\epsilon}) = \text{argmax}_{c^1,\ldots,c^k,\epsilon} \prod_{s=1}^{n} \max_{h\in\{1,\ldots,k\}} f^{c^h,\epsilon}(x_1^s, \ldots, x_{t_s}^s) \tag{4.1}$$

Now, we have defined for each allowable number of decision rules k how we can estimate the maximum likelihood collection of rules. By defining

$$\hat{I}_s^h = \begin{cases} 1 & \text{if } h = \text{argmax}_{h'\in\{1,\ldots,k\}} f^{\hat{c}^{h'},\epsilon}(x_1^s, \ldots, x_{t_s}^s); \\ 0 & \text{otherwise} \end{cases}$$

Then, we can estimate p_h, the proportion of our population using rule h, via the maximum likelihood estimate

$$\hat{p}_h = \frac{1}{n}\sum_{s=1}^{n} \hat{I}_s^h.$$

In the following section, we shall prove the consistency of our estimates of k, $(c^1, \ldots, c^k)$, and $(p_1, \ldots, p_k)$. But in order to get a consistent estimate of k, we need first to introduce an information criterion which incorporates a penalty for the number of parameters in the model.[1] We shall prove that any strictly increasing penalty function with bounded increments will yield consistent estimate of k, the c^h's, ϵ, and the p_h's (all we actually need is that the increments of the penalty not increase at too high a rate); but we propose one particular such penalty with a pseudo-Bayesian interpretation. If we start with a prior α_k on the model with k rules, and for each such model introduce a prior $\pi_k(c^1, \ldots, c^k) \otimes \mu_k(d\epsilon)$, then the Bayes procedure will pick the model which maximizes the log Bayes posterior

$$B(X,k) = \log \int_0^1 \sum_{c_1,\ldots,c^k} \alpha_k \pi_k(c^1, \ldots c^k) \prod_{s=1}^{n} \max_{h\in\{1,\ldots,k\}} f^{c^h,\epsilon}(x_1^s, \ldots, x_{t_s}^s)\mu_k(d\epsilon)$$

Choosing the prior $\pi_k(.)$ to be uninformative (assigning prior probability $\frac{1}{8^{3k}}$ to each possible k-tuple of rules in $\mathfrak{C}^k$), letting $\mu_k(d\epsilon)$ be uniform for all k, letting $\alpha_k = \frac{1}{2^k}$, and assuming that the maximal likelihood term dominates all other terms in the Bayes posterior (which assumption is asymptotically justified), we choose the model which maximizes

$$\log\Big(\prod_{s=1}^{n} \max_{h\in\{1,\ldots,k\}} f^{\hat{c}^h,\hat{\epsilon}}(x_1^s, \ldots, x_{t_s}^s)\Big) - 3k\log(8) - k\log(2).$$

[1] There is a substantial literature on the problem of choosing an optimal penalty for the complexity of a model. The best known, and one of the earliest, is Akaike (1974)'s criterion (which picks the model that maximizes the maximal log likelihood less the number of parameters). Another very popular information criterion was introduced by Schwarz (1978), which picks the model that maximizes log of the maximal likelihood less the number of parameters multiplied by log of the sample size, divided by two. Many other criteria are implicit in the coding literature such as Wallace and Boulton (1968)'s Minimum Message Length, and Rissanen (1978)'s Minimum Description Length. Each of these procedures has its epistemic advantages, and some (e.g. Schwarz (1978)) have known asymptotic properties for a given class of likelihood functions.

Notice that if we take logs of base 2, and if our parameter space $\mathfrak{C}^k$ were continuous, this would reduce to the Akaike information criterion. We are now ready to prove the consistency of the general class of procedures to which our procedure belongs. The readers who are not interested in the proof of the consistency of our procedure can skip the following section and proceed directly to section 6.

5 Consistency of Our Procedure

We start this section by a more general statement of the procedure. We collect data on n individuals, each of whom gets to perform t independent tasks. The number of tasks can be different for different individuals, but since we study the asymptotics as $t \uparrow \infty$, and then as $n \uparrow \infty$, we shall simplify the notation by suppressing that possible difference. There are k rules being used in the population, each defining a likelihood function $f_h \in \mathfrak{F}$ (which is a density function on $\mathbb{R}$). Without loss of generality, we further assume that the k rules are distinct; i.e. if $l \neq h$, then $f_l \neq f_h$. Of course, if two rules were observationally equivalent (induce the same likelihood function), then there is no loss in treating them as one rule. Our data is of the form $\{x_j^i; i = 1, \ldots, n; j = 1, \ldots, t\}$. Let $I_i^h = 1$ if agent i uses rule h, and $I_i^h = 0$ otherwise. We further assume that $x_1^i, \ldots x_t^i$ is an i.i.d. sample from f_h if $I_i^h = 1$, and that $x_j^i \in \mathbb{R}; i = 1, \ldots, n; j = 1, \ldots, t$.[2] We assume that in the population from which we sample, the proportion using rule h is p_h, where $-1 \leq p_h \leq 1$, and $\sum_{j=1}^{k} p_h = 1$.

We can now define our estimates $\hat{k}, \hat{f}_1, \ldots, \hat{f}_{\hat{k}}, \hat{p}_1, \ldots, \hat{p}_{\hat{k}}$ as follows.

$$(\hat{k}, \hat{f}_1, \ldots, \hat{f}_{\hat{k}}) = \text{argmax}_{k', f_1, \ldots, f_{k'}} \left\{ \sum_{i=1}^{n} \left(\max_{h \in \{1, \ldots, k'\}} \sum_{j=1}^{t} \log f_h(x_j^i) \right) - penalty(k') \right\}.$$

Where $penalty(k)$ is some deterministic increasing function of k, with $m \leq penalty(k) - penalty(k-1) \leq M$ for some finite positive numbers m, M. We then estimate

$$\hat{p}_h = \frac{1}{n} \sum_{i=1}^{n} \hat{I}_i^h,$$

where

$$\hat{I}_i^h = \begin{cases} 1 & \text{if } h = \text{argmax}_{l \in \{1, \ldots, \hat{k}\}} \sum_{j=1}^{t} \log \hat{f}_l(x_j^i); \\ 0 & \text{otherwise} \end{cases}$$

Theorem 1

$$\Pr\left\{ \lim_{n \uparrow \infty} \lim_{t \uparrow \infty} (\hat{k}, \hat{f}_1, \ldots, \hat{f}_{\hat{k}}, \hat{p}_1, \ldots, \hat{p}_{\hat{k}}) = (k, f_1, \ldots, f_k, p_1, \ldots, p_k) \right\} = 1.$$

[2] For instance, in our case, the sufficient statistic for each observation x_τ^s is the number $x_{c,\tau}^s \in \{0, 1\}$. The results below will still hold if we allow each x_j^i to be a vector in $\mathbb{R}^d$, but in most cases, a real sufficient statistic will suffice.

Proof: We start by considering a given estimate $\hat{k}$, and the corresponding sequence of estimates $\hat{f}_1, \ldots, \hat{f}_{\hat{k}}$, as $t \uparrow \infty$. For all subjects i with $I_i^h = 1$, define the empirical distribution function

$$\tilde{F}_i^t(\xi) = \frac{1}{t}\#\{x_\tau^i | x_\tau^i \leq \xi\}.$$

Under the null hypothesis, and since the f_h's are assumed distinct, by the Glivenko-Cantelli Lemma, letting $F_h(\xi) = \int_{-\infty}^{\xi} f_j(x)\, dx$, we have

$$\Pr\left\{\sup_{\xi \in \mathbb{R}} |\tilde{F}_i^t(\xi) - F_h(\xi)| \to 0\right\} = \begin{cases} 1 & \text{if } I_i^h = 1; \\ 0 & \text{otherwise.} \end{cases}$$

Hence, given any two individuals i, i' such that $I_i^h = I_{i'}^h = 1$, as $t \uparrow \infty$, their two empirical distributions converge almost surely, which in turn almost surely guarantees that they will be classified to the same rule, as $t \uparrow \infty$. Therefore, for each $\gamma > 0$, there exists a sufficiently large T such that $\forall t \geq T$, if i, i' belong to the same class, for all $1 \leq h \leq \hat{k}$, $\Pr\{|\hat{I}_i^h - \hat{I}_{i'}^h| < \gamma\} = 1$. Choosing $\gamma < \frac{1}{2}\min_{h'}(max_\xi |\log f_h(\xi) - \log f_{h'}(\xi))|$, all subjects of class h will be classified together with probability one for all $t \geq T$.

Now, for each pair i, i' that belong to the same class, let $T_{i,i'}$ be the number of tasks needed to have the two in the same class with probability one (such a $T_{i,i'}$ exists by the previous paragraph). Let $T^\star = \max_{\{1 \leq i, i' \leq n\}} T_{i,i'}$. Then for all $t \geq T^\star$, all individuals belonging to the same group will have been allocated together with probability one. Now, if $\hat{k} < k$, that means that some classes have been grouped together. Let n' be a group of individuals who use rule h, and who are combined with another group that uses rule h'. Let the estimated rule for the combined group be $f_{\hat{h}}$. The gain from adding one more rule is greater than $\sum_{i=1}^{n'} \sum_{j=1}^{t} [log(f_h(x_j^i)) - log(f_{\hat{h}}(x_j^i))]$. For all $t \geq T$, and as t gets larger, this term diverges, eventually exceeding M (the bound on the incremental penalty for adding an extra rule). Hence, for all $t \geq T^\star$, the probability of $\hat{k} < k$ is zero.[3] On the other hand, if $\hat{k} > k$, then as $t \uparrow \infty$, $\hat{k} - k$ rules will have no members (since all members of the same class are classified

[3] In our application, let $\hat{h}$ (the rule to which the n' subjects are classified) differ from h (their true rule) in a collection of observations (prior and number of drawn N's) that occur with some probability p (notice that $p > 0$ by the distinctness of the rules). In the limit the contribution of each of the subjects when classified to rule $\hat{h}$ when indeed they use rule h is

$$\left(1 - \frac{\epsilon}{2}\right)^{(1-p)t(1-\frac{\epsilon}{2})} \times \left(\frac{\epsilon}{2}\right)^{(1-p)t\frac{\epsilon}{2}} \times \left(\frac{\epsilon}{2}\right)^{pt(1-\frac{\epsilon}{2})} \times \left(1 - \frac{\epsilon}{2}\right)^{pt\frac{\epsilon}{2}},$$

whereas their contribution to the likelihood function when they are correctly classified to rule h is

$$\left(1 - \frac{\epsilon}{2}\right)^{t(1-\frac{\epsilon}{2})} \times \left(\frac{\epsilon}{2}\right)^{t\frac{\epsilon}{2}}.$$

The likelihood ratio (the second term divided by the first, measuring the gain from creating a new rule for that one individual) is $\left(\frac{1-\frac{\epsilon}{2}}{\frac{\epsilon}{2}}\right)^{pt(1-\epsilon)}$, which for $\epsilon < 1$ diverges to infinity as $t \uparrow \infty$, thus proving that the log likelihood gain will eventually overtake the maximum penalty M, and a new rule will be allowed.

together and there are only k classes), hence the contribution to the likelihood from those extra classes will converge to zero, eventually falling below m, and we shall discard those extra rules. Therefore, for all $t \geq T^\star$, the probability of $\hat{k} > k$ is zero. Combining the two results, we know that $\hat{k} = k$ almost surely, for all $t \geq T^\star$.

Now that we have established that as $t \uparrow \infty$, the number of classes k is almost surely correctly estimated by $\hat{k}$, and the individuals are almost surely correctly classified with the other individuals who use the same rule, the rest of the results follow trivially. The almost sure estimation of $p_1, \ldots, p_k$ correctly by $\hat{p}_1, \ldots, \hat{p}_{\hat{k}}$ as $n \uparrow \infty$ follows from the strong law of large numbers (moreover, a central limit theorem applies, and we can construct confidence intervals for the p_h's). The consistent estimation of the rules $f_1, \ldots, f_k$ also follows immediately from the Glivenko-Cantelli Lemma (and pointwise Central limit theorems apply, which we can use to estimate confidence intervals for the f_h's). ∎

6 The Data Analysis

Now, having established the consistency of our estimation/classification procedure, we proceed to discuss its implementation. For a given k, a rather crude method for estimating $c^1, \ldots, c^k$ and ϵ would be to loop over all possible classifications of the n subjects to the various rules, estimate the maximal likelihood rules used by each class, and pick the classification that yields the maximal overall likelihood. This brute-force method would not be feasible for reasonably large k and n, since for each k-tuple of rules, the number of possible classifications is k^n (i.e. it grows geometrically in k, and exponentially in n, since we want n to be very large, this procedure would be computationally infeasible). Fortunately, there is no need to explicitly loop over all possible classifications. We can implement the classification implicitly by using a standard maximization routine for the likelihood function with k rules (the dimensionality of the parameter space is now $3k$, and the number of possible k-tuples of rules is 8^{3k}). For each likelihood evaluation of the routine, as we loop through the data of each of the n individuals, we make k likelihood evaluations for that individual's data using each of the k rules. For the given class of rules under consideration, we classify each subject to the rule that maximizes the likelihood of that subject's data, and the contribution of that subject's data to the overall log likelihood function is added to the other subjects' as shown in equation (4.1). Notice that this transforms our problem into a simple likelihood maximization one, with the number of parameters growing linearly in the number of rules, and each likelihood function evaluation growing linearly in the number of agents and number of rules. This improvement over the brute-force (geometric/exponential) algorithm makes our procedure rather easy to implement by invoking any of the standard multidimensional optimization subroutines generally available in mathematical and statistical packages.

The results that we obtained by applying our algorithm to the data from the four universities are reported in the seven tables in Appendix A. We had a total of 256 subjects, and the total number of tasks was 4506. The first 4 tables report the results where we analyze the data for each of the schools separately. Table I shows our analysis the data from UCLA, Table II shows the analysis of the data from PCC, Table III shows the analysis of the data from

Occidental College, and Table IV shows the analysis of the data from CSULA. Tables I-IV report estimates up to $k = 7$. Tables V and VI report the maximimum likelihood estimates, and the value of our information criterion, for all the subjects who were payed according to the outcome, and all the subjects who were payed a flat fee, respectively. Table VII reports the estimates and information criterion for $k = 1, 2, 3$, for all the subjects in our sample pooled together. In each of these tables, for a number of given k's, we report the maximum likelihood estimates of ϵ, the maximum likelihood estimate of $(c_1^1, c_2^1, c_3^1), \ldots, (c_1^k, c_2^k, c_3^k)$ (with the number of subjects allocated to each of the rules shown in parentheses), the value of the maximal likelihood for the model with k rules, and the information criterion that we introduced in Section 4 (IC = log(maximal likelihood) - $3k$ log(8) - k log(2)). When our information criterion told us to stop after a certain number of rules, we indicated that by italicizing the values of the IC that tell us that we should not use a model with that many parameters.

In order to facilitate the understanding of the rules that our algorithm estimated, we now identify three of those rules:

- A subject who correctly uses Bayes's rule chooses cage A if the prior in favor of cage A was 1/3 and the number of N's was greater than 4, the prior in favor of A was 1/2 and the number of N's was greater than 3, or the prior in favor of A was 2/3 and the number N's observed was more than 2; otherwise cage B is chosen. In our notation, that means that the cutoff rule (c_1, c_2, c_3)=(432) corresponds to Bayes's rule.

- A subject who judges likelihood using the representativeness heuristic would have the same cutoff for all three priors (i.e. $c_1 = c_2 = c_3$). Since 3 is representative of the parent distribution in cage B, and 4 is representative of the parent distribution in cage A, a subject using the representativeness heuristic will pick cage B if 3 N's are drawn, and pick cage A if 4 N's are drawn. The representative heuristic together with our rationality assumption that more N's should be treated as increased evidence in favor of cage A, makes a subject using that heuristic pick cage B if the number of observed N's is less than or equal to 3, and cage A if the numbe of observed N's is greater than or equal to 4, resulting in the cutoff rule (333).

- A third class of subjects that we wish to identify are conservative Bayesians. Those subjects give more weight to the prior odds than Bayes's formula (posterior odds = prior odds $\times$ likelihood ratio) dictates. For instance, subjects using the cutoff rule (531), requiring six N's to be observed in order to pick cage A (where five N's would be enough to convince a Bayesian to pick that cage) when the prior favoring A is 1/3, and one N to be observed ~~in order to pick cage B (where two N's~~ would be enough to convince a Bayesian to make that choice) when the prior favoring A is 2/3, are clearly giving too much weight to the prior information and hence needing extra evidence to be convinced to change their prior pick. Note that due to the discreteness of our observations, subjects could be conservative and yet use the rule (432). However, subjects using (531) are definitely conservative (see Edwards (1982)).

We summarize the most important aspects of our results below:

- For all tables but Occidental College and CSULA, when we force the algorithm to choose only one rule, it picks the rule (432), which corresponds to Bayesian updating. Even in the two institutions where (432) was not picked initially as the first most likely rule, when we allowed the algorithm to pick more rules, (432) surfaced and dominated all the other rules. With the exception of the PCC table, the (432) Bayesian rule is the most prominent (with more subjects allocated to it than any other rule) regardless of how many rules we allowed the algorithm to pick.

- The second most prominent rule in all but the PCC table (where it is the most prominent) is (333) which corresponds to making the decision based on the likelihood ratio, ignoring the prior. This rule has been named the "representativeness" rule by psychologists, and its robustness regardless of the number of rules that we allow our algorithm to pick is also remarkable.

- The third most prominent rule that we pick in all tables once enough rules are allowed is (531) which corresponds to weighting the prior more than its share in Bayes's formula (posterior odds = prior odds $\times$ likelihood ratio). This rule also has been discussed in the psychology literature, and it has been given the name "conservatism" since it requires more data to change one's mind than Bayes's rule dictates.

- We note that the algorithm was free to pick any of the possible rules. For instance, when three rules are allowed, where most of the Tables have picked the rules (432), (333), and (531) in that order, there are a total of $8^9 = 134,217,728$ rules available, of which the algorithm picked our three most prominent rules.

- Of our four schools, UCLA had the lowest estimated ϵ, followed by PCC, followed by Occidental College, and CSULA had the highest estimate of the error rate ϵ. Also, Tables V and VI show that our estimate of ϵ for the subjects who were payed according to the outcome was lower than its counterpart for the subjects who were payed a flat fee. Moreover, the ordering of the schools, and of those who were payed according to the two systems, is almost the same when we inspect the proportions who use Bayes's rule.

Some cautionary notes are in order before we conclude this paper. Even though our procedure has been proven to yield consistent estimates of the number of rules, and the rules themselves, there are a few issues that may arise in finite samples. We discuss some of those issues in the framework of our application.

- In finite samples, certain individuals' data could have the same likelihood under two or more rules. For instance, it is conceivable that an individual's collection of observations (p, N) (where p is the induced prior, and N is the number N's observed) all fall in the class of observations that are treated identically by rules (531) and (521) (for instance, if there were no data points where prior 2 was induced, and 3 N's were drawn). In those cases, the likelihood function will have multiple maxima, and we are free to choose any one of them. In the tables reported in the appendix, we gave the benefit of the doubt to rules that were a priori more appealing to us (e.g. (432) was chosen any time it was tied with one of the other rules). This limitation is not severe since in large samples, the multiple maxima will vanish, and since the rules we gave priority in breaking ties were the ones that were picked by the algorithm when fewer rules were allowed for (e.g. where (432) is the rule being picked when only one is allowed).

- Our information criterion is not sufficiently severe in punishing models with a large number of rules. For instance, where it did indicate that we should stop in the UCLA Table I with six rules, one rule has only 4 subjects (less than 5% of the population) allocated to it. When we allowed for a seventh rule, only 2 subjects were allocated to that rule, and now the information criterion's penalty for the extra rule was enough to tell us to stop. The worst illustration of the permissiveness of our information criterion is in the Occidental College Table III, where the seventh rule (111) (which is a very strange rule, equivalent to almost always choosing cage A, regardless of the prior and outcome) only picked one individual out of 56, and our information criterion (which told us to stop with five rules) does not prefer six rules to seven. Of course, on the positive side, the rule of almost always picking cage A may be a legitimate rule that some subjects are using. Our theory suggests that if such a rule were correct, regardless of how few individuals used it, then it should indeed be picked by the algorithm. If it is not a rule that anyone uses, but we are simply picking the idiosyncratic behavior of a subject in this particular collection of tasks, then more tasks should be assigned to that subject in order to be able to determine if that is a legitimate rule.

We note on the positive side that those issues arising in finite samples can be resolved not only by letting the number of tasks and the number of individuals get large, but also by considerations of optimal experimental design. Our data set was not constructed for the purposes of this paper (see Grether (1980) for discussion). For instance, if we were interested in distinguishing models which give different weights to the prior odds and the likelihood ratio (which seems a reasonable goal after we saw the rules that our algorithm picked), then the design with the flat prior (prior 2) would probably have been omitted due to its uninformativeness. Also, the data generating mechanism was chosen to increase the probability of getting outcomes that mimic the parent distribution. That design may be suboptimal for our purposes, and we probably would have chosen a larger number of priors, and more draws from the cages to reduce the probability of ties between various rules. The essential point to note is that our procedure's performance can be made as good as one wishes, even in finite samples (given t and n), by studying issues of optimal experimental design (choice of priors, and the number of draws from each cage, for instance) to minimize the occurrence of ties, and maximize the sharpness of our information criterion (which serves as an Ockham's razor).

8 Appendix A

Table I				
UCLA				
No. of Rules	ϵ	Rules (#)	Likelihood	IC
1	0.3	432 (97)	6.726×10^{-363}	-840.864
2	0.3	432 (71) 333 (26)	3.026×10^{-328}	-768.003
3	0.2	432 (50) 333 (26) 531 (21)	7.018×10^{-311}	-734.950
4	0.2	432 (46) 333 (24) 531 (19) 443 (8)	2.447×10^{-301}	-719.909
5	0.2	432 (46) 333 (24) 531 (17) 443 (4) 543 (6)	1.300×10^{-295}	-713.658
6	0.2	432 (46) 333 (19) 531 (17) 443 (6) 311 (5) 541 (4)	7.769×10^{-291}	-709.591
7	0.2	432 (44) 333 (19) 531 (17) 443 (6) 311 (5) 541 (4) 441 (2)	6.220×10^{-289}	*-712.139*

Table II				
PCC				
No. of Rules	ϵ	Rules (#)	Likelihood	IC
1	0.4	432 (66)	7.412×10^{-203}	-472.353
2	0.3	432 (31) 333 (35)	2.276×10^{-162}	-386.059
3	0.3	432 (25) 333 (35) 531 (6)	7.770×10^{-155}	-375.645
4	0.2	432 (20) 333 (33) 531 (6) 444 (7)	5.668×10^{-148}	-366.774
5	0.2	432 (18) 333 (28) 531 (6) 444 (7) 332 (7)	3.012×10^{-142}	-360.522
6	0.2	432 (15) 333 (23) 531 (6) 444 (7) 332 (7) 433 (8)	1.601×10^{-136}	-354.270
7	0.2	432 (13) 333 (23) 531 (6) 444 (5) 332 (7) 433 (8) 111 (0)	1.601×10^{-136}	*-361.201*

Table III

Occidental College

No. of Rules	ϵ	Rules (#)	Likelihood	IC
1	0.4	332 (56)	4.075×10^{-206}	-479.859
2	0.3	332 (47) 511 (9)	2.831×10^{-185}	-438.801
3	0.3	432 (28) 333 (19) 511 (9)	3.300×10^{-170}	-411.040
4	0.3	432 (21) 333 (13) 332 (14) 511 (8)	1.988×10^{-163}	-402.360
5	0.3	432 (16) 333 (13) 532 (8) 332 (12) 511 (7)	1.616×10^{-159}	-400.288
6	0.3	432 (18) 333 (12) 332 (14) 542 (3) 511 (7) 222 (2)	1.198×10^{-156}	-400.611
7	0.2	432 (19) 333 (14) 332 (10) 531 (4) 542 (3) 511 (5) 111 (1)	1.976×10^{-153}	-400.135

Table IV				
CSULA				
No. of Rules	ϵ	Rules (#)	Likelihood	IC
1	0.5	433 (37)	1.141×10^{-169}	-395.936
2	0.4	432 (24) 333 (13)	6.137×10^{-152}	-362.042
3	0.4	432 (16) 333 (12) 433 (9)	4.022×10^{-147}	-357.883
4	0.4	432 (11) 333 (11) 531 (7) 433 (8)	1.748×10^{-143}	-356.437
5	0.3	432 (10) 333 (11) 531 (6) 542 (2) 433 (8)	9.598×10^{-141}	*-357.060*
6	0.3	432 (9) 333 (11) 531 (5) 433 (7) 542 (2) 444 (3)	3.082×10^{-139}	*-360.523*
7	0.3	432 (8) 333 (11) 531 (5) 433 (7) 444 (3) 542 (2) 442 (1)	9.897×10^{-138}	*-363.985*

Table V				
All Schools–payed according to outcome				
No. of Rules	ϵ	Rules (#)	Likelihood	IC
1	0.4	432 (125)	1.229×10^{-431}	-999.140
2	0.3	432 (85) 333 (40)	1.354×10^{-382}	-893.147
3	0.3	432 (64) 333 (38) 531 (23)	1.678×10^{-361}	-851.510
4	0.3	432 (55) 333 (31) 531 (22) 433 (17)	1.042×10^{-351}	-835.892
5	0.3	432 (44) 333 (28) 531 (21) 433 (20) 332 (12)	4.595×10^{-344}	-825.222
6	0.2	432 (43) 333 (26) 531 (19) 433 (19) 332 (12) 543 (6)	2.198×10^{-337}	-816.773

Table VI				
All Schools–payed a flat fee				
No. of Rules	ϵ	Rules (#)	Likelihood	IC
1	0.4	432 (131)	1.188×10^{-520}	-1204.104
2	0.3	432 (77) 333 (54)	3.693×10^{-449}	-1045.417
3	0.3	432 (56) 333 (53) 521 (22)	2.594×10^{-427}	-1003.045
4	0.3	432 (51) 333 (51) 542 (10) 521 (19)	5.174×10^{-416}	-983.958
5	0.3	432 (49) 333 (40) 531 (10) 433 (19) 332 (13)	1.767×10^{-408}	-973.543
6	0.3	432 (41) 333 (38) 542 (10) 521 (19) 332 (13) 433 (10)	1.064×10^{-401}	-964.864

Table VII				
All Schools				
No. of Rules	ϵ	Rules (#)	Likelihood	IC
1	0.4	432 (256)	1.460×10^{-951}	-2196.312
2	0.3	432 (162) 333 (94)	4.999×10^{-831}	-1925.702
3	0.3	432 (120) 333 (92) 531 (44)	1.355×10^{-789}	-1837.230

9 Appendix B

Table 1: Raw data on 20 tasks for 48 subjects in UCLA who were each payed a bonus of $10 if they made the correct choice on a randomly drawn task. The first line contains a list of the 20 priors for the 20 tasks (2 refers to the prior of 1/3 in favor of cage A, 3 refers to the prior of 1/2 in favor of cage A, and 4 refers to the prior of 2/3 in favor of cage A). The second line contains a list of the 20 outcomes (number of N's observed, between 0 and 6). The last 48 lines contain the 20 choices of each of the subjects (choosing cage A is coded as 1, and choosing cage B is coded as 0), for each of the tasks.

2	2	4	4	3	3	2	2	4	4	3	3	2	2	4	4	3	3	2	2
0	4	4	4	4	2	4	3	4	5	4	5	5	4	4	4	5	5	4	4
0	1	1	1	1	0	1	0	1	1	1	1	1	0	1	1	1	1	1	0
0	0	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	1	0	0
0	1	1	1	0	0	1	0	1	1	1	1	0	0	1	1	1	0	0	0
0	1	1	1	1	0	1	0	1	1	1	1	1	0	1	1	1	1	0	0
0	1	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	0
0	1	1	1	0	1	0	0	1	1	1	1	1	0	1	1	1	1	0	0
0	0	1	0	1	0	0	0	1	1	0	1	1	1	1	1	1	1	0	0
0	0	1	1	1	0	0	0	1	1	1	1	1	0	1	1	1	1	0	0
0	1	1	1	1	0	1	0	1	0	1	1	1	0	1	0	1	1	0	0
0	0	1	1	1	0	0	0	1	1	1	1	1	0	1	1	1	1	0	0
0	0	1	1	0	0	0	0	1	1	0	1	0	0	1	1	1	1	0	0
0	1	1	1	1	0	1	0	1	1	0	1	1	0	1	0	0	1	0	0
0	1	1	1	0	0	0	0	1	1	1	1	1	0	1	1	1	1	0	0
0	0	1	1	1	0	0	0	1	1	1	1	1	0	1	1	1	1	0	0
0	0	1	1	1	0	0	0	1	1	1	1	1	0	1	1	1	1	0	0
0	0	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	0	0
0	1	0	1	1	0	0	0	1	1	0	1	1	0	1	1	1	1	0	0
0	0	1	1	1	0	1	0	1	1	1	1	1	0	1	1	1	1	0	0
0	0	1	1	1	0	0	0	1	1	1	1	0	0	1	1	1	1	0	0
0	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	0	0
0	0	1	1	1	0	0	0	1	1	1	1	1	0	1	1	1	1	0	0
0	0	1	1	0	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0
0	0	1	1	1	0	0	0	1	1	1	1	0	0	1	1	1	1	0	0
0	0	1	1	1	0	0	0	1	1	1	1	0	0	1	1	1	1	0	0
0	1	1	1	1	0	0	0	1	1	1	1	1	0	1	1	1	1	0	0
0	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	0	0
0	1	1	1	1	0	1	0	1	1	1	1	0	1	1	1	1	1	0	0
0	1	1	1	1	0	1	0	1	0	1	1	0	0	0	1	1	1	0	1
0	0	0	1	0	0	0	0	1	1	1	1	1	0	0	1	1	0	0	0
0	0	1	1	1	0	0	0	0	1	1	0	0	0	1	1	1	1	0	1
0	1	1	1	1	0	0	1	1	1	1	1	1	0	1	0	1	1	1	0
0	0	1	1	1	0	0	0	1	1	1	1	0	0	1	1	1	1	0	0
0	0	1	1	1	0	0	0	1	1	0	1	1	0	1	1	1	1	0	1
1	1	0	1	0	0	0	1	0	1	1	0	0	0	1	1	1	1	0	1
0	0	1	1	1	0	0	0	1	1	1	1	1	0	1	1	1	1	0	0
0	0	1	1	0	0	0	0	1	1	0	1	1	0	1	1	1	1	0	0
0	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0
0	1	1	1	1	0	1	0	1	1	0	1	1	0	1	1	1	1	0	0
0	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1
0	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1
0	1	1	0	0	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1
0	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1
0	1	1	1	1	0	1	0	1	1	0	1	1	1	1	1	1	1	0	0
0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	0	1	0
0	1	1	1	1	0	0	0	1	1	0	1	1	0	1	0	1	1	0	0
0	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	0	0
0	1	1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1	1	0
0	1	0	1	1	0	0	0	1	1	1	1	1	0	0	1	1	1	0	0

Table 2: Raw data on 20 tasks for 49 subjects in UCLA who were each payed a flat fee for participating in the experiment. The first line contains a list of the 20 priors for the 20 tasks (2 refers to the prior of 1/3 in favor of cage A, 3 refers to the prior of 1/2 in favor of cage A, and 4 refers to the prior of 2/3 in favor of cage A). The second line contains a list of the 20 outcomes (number of N's observed, between 0 and 6). The last 49 lines contain the 20 choices of each of the subjects (choosing cage A is coded as 1, and choosing cage B is coded as 0), for each of the tasks.

2	2	4	4	2	4	4	4	3	3	4	4	4	2	4	3	4	2	4	4
4	4	1	4	3	4	3	6	4	4	5	5	5	4	2	2	4	3	4	3
0	1	0	1	0	1	1	1	0	1	1	1	1	0	0	0	1	0	1	1
1	0	0	1	0	1	0	1	1	0	1	1	1	0	0	0	1	0	1	0
0	0	0	1	0	1	0	1	0	0	1	1	1	1	0	0	1	0	1	0
0	1	0	1	0	1	1	1	0	1	1	1	0	0	0	0	1	1	1	1
1	1	0	1	0	1	0	1	1	1	1	1	1	0	0	0	1	0	1	1
0	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	1	0	1	1
1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1
0	0	0	1	0	1	0	1	1	0	1	1	1	0	0	0	1	0	0	0
1	1	0	0	0	1	1	1	1	1	1	1	1	0	0	0	1	0	1	1
1	0	0	1	0	1	0	1	0	1	1	0	1	1	1	1	1	0	1	0
1	1	0	1	0	1	0	1	1	1	1	1	1	0	0	0	1	0	1	1
0	1	0	1	0	1	0	1	1	0	1	1	1	1	0	0	1	0	1	0
0	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	1	0	1	1
0	0	0	0	0	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1
0	1	0	1	0	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1
1	1	0	1	0	1	0	1	1	1	1	1	1	0	0	0	1	0	1	1
0	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	1	0	1	1
0	0	0	1	0	1	1	1	0	0	1	1	1	0	1	1	1	0	1	1
1	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	1	0	1	1
0	1	0	1	0	1	0	1	1	0	1	1	0	0	0	0	1	0	1	1
1	1	0	1	0	1	0	1	1	1	0	1	1	0	0	1	1	0	1	0
1	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	1	0	1	0
0	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	1	0	1	1
1	0	0	1	0	0	1	1	0	1	1	1	1	0	0	0	1	0	1	0
1	1	0	1	0	1	1	1	0	1	1	1	1	1	0	0	1	0	1	1
0	0	0	1	0	0	1	1	1	1	1	1	1	0	1	0	1	0	1	1
0	0	1	1	0	1	1	1	1	1	1	1	1	0	0	0	1	0	1	1
1	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	1	0	1	1
1	0	0	1	0	1	1	1	1	0	1	1	1	0	0	0	1	0	1	0
1	1	0	1	0	1	1	1	1	0	1	1	1	0	0	0	1	0	1	0
1	0	1	1	0	1	0	1	0	0	0	1	1	0	1	0	1	0	1	1
0	0	0	1	0	1	0	1	1	1	1	1	1	0	0	0	1	0	1	1
0	1	1	0	1	0	1	1	1	1	0	1	1	0	0	0	1	0	1	1
0	0	0	1	0	1	1	1	1	1	1	1	1	1	0	0	1	0	1	0
1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0
0	0	0	1	0	1	0	1	1	1	1	1	1	0	0	0	1	0	0	0
0	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	1	0	1	0
0	1	0	1	0	1	1	1	1	1	1	1	1	0	0	0	1	0	1	1
0	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	1	0	1	1
1	0	0	1	0	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1
1	1	1	1	0	0	0	0	1	1	1	1	1	0	0	0	1	0	1	1
0	1	0	1	0	1	1	1	1	0	1	1	1	1	0	0	1	0	1	0
1	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	1	0	1	0
0	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1
0	1	0	1	0	1	0	1	1	1	1	1	1	0	0	0	1	0	1	0
1	1	1	1	0	1	1	1	0	0	1	1	1	0	1	0	1	0	1	1
1	0	0	1	0	1	0	1	1	1	1	1	1	1	0	0	1	0	1	0
0	0	1	1	0	1	1	1	1	1	1	1	1	0	0	0	1	1	1	0
1	0	0	1	0	1	1	1	1	1	1	1	1	0	1	0	1	0	1	0

Table 3: Raw data on 14 tasks for 35 subjects in PCC who were each payed a bonus of $10 if they made the correct choice on a randomly drawn task. The first line contains a list of the 14 priors for the 14 tasks (2 refers to the prior of 1/3 in favor of cage A, 3 refers to the prior of 1/2 in favor of cage A, and 4 refers to the prior of 2/3 in favor of cage A). The second line contains a list of the 14 outcomes (number of N's observed, between 0 and 6). The last 35 lines contain the 14 choices of each of the subjects (choosing cage A is coded as 1, and choosing cage B is coded as 0), for each of the tasks.

2	2	4	4	2	2	4	4	3	3	2	2	4	4
4	2	3	2	5	5	4	4	3	3	2	3	3	6
1	0	0	0	1	1	1	1	0	0	0	0	1	1
0	0	1	0	1	1	1	1	0	0	0	0	1	1
0	0	1	1	0	0	1	0	0	0	0	0	1	1
0	0	1	0	1	1	1	1	0	0	0	0	1	1
0	0	1	0	1	1	1	1	0	0	0	0	1	1
0	0	1	0	1	1	1	1	0	0	0	1	0	1
0	0	1	1	0	0	1	1	0	1	0	0	1	1
0	1	1	0	1	1	1	0	0	0	0	0	0	1
0	0	0	0	1	1	1	1	1	0	0	0	0	1
0	0	1	0	1	1	1	1	0	0	0	0	1	1
1	0	0	0	1	1	0	1	0	0	0	0	1	1
1	0	0	0	1	1	1	1	0	0	0	0	0	1
1	0	0	0	1	1	1	1	0	0	0	0	0	1
1	0	0	0	1	1	1	1	0	0	0	0	0	1
0	0	0	0	1	1	1	1	0	0	0	0	0	1
0	0	1	0	1	1	0	0	0	0	0	0	1	1
0	1	0	0	1	1	1	1	0	0	0	1	1	1
0	0	0	0	1	1	1	1	0	1	0	1	1	1
1	0	0	0	1	1	0	1	0	0	0	0	0	1
0	0	0	0	0	1	1	1	0	0	0	0	0	1
1	0	1	0	1	1	1	1	0	0	0	1	1	1
0	0	1	0	0	1	1	0	1	0	0	0	0	1
0	0	0	0	1	1	0	0	0	0	0	0	0	1
0	1	0	0	1	1	1	0	1	1	0	0	1	1
0	0	0	1	1	1	1	1	0	0	0	0	0	1
1	0	0	0	1	1	1	1	0	0	0	0	1	1
0	0	1	0	1	0	0	1	0	1	1	0	1	0
0	0	0	0	0	1	1	0	0	0	0	1	1	1
0	0	0	1	1	1	1	0	0	0	0	0	1	1
1	0	1	0	1	1	0	0	0	0	0	0	0	1
0	0	1	0	0	1	0	1	0	0	0	0	1	1
0	0	0	0	1	1	1	1	0	0	0	1	0	1
1	0	0	0	1	1	1	0	1	0	0	0	1	1
1	0	1	0	1	1	1	1	0	0	0	0	1	0
1	0	1	1	1	1	1	0	0	0	0	0	0	1

Table 5: Raw data on 14 tasks for 25 subjects in Occidental College who were each payed a bonus of $10 if they made the correct choice on a randomly drawn task. The first line contains a list of the 14 priors for the 14 tasks (2 refers to the prior of 1/3 in favor of cage A, 3 refers to the prior of 1/2 in favor of cage A, and 4 refers to the prior of 2/3 in favor of cage A). The second line contains a list of the 14 outcomes (number of N's observed, between 0 and 6). The last 25 lines contain the 14 choices of each of the subjects (choosing cage A is coded as 1, and choosing cage B is coded as 0), for each of the tasks.

2	2	4	4	2	2	4	4	3	3	2	2	4	2
1	4	4	2	3	3	5	3	4	3	5	3	3	3
0	1	1	0	0	0	1	0	1	0	1	0	1	0
0	0	1	0	0	0	1	0	1	0	1	0	0	0
0	1	1	0	0	0	1	1	1	0	0	0	1	0
0	0	1	0	0	0	1	1	1	0	1	0	1	0
0	1	0	0	1	0	1	1	0	0	1	0	1	0
0	1	1	0	0	0	1	1	1	0	1	0	1	0
0	1	1	1	0	0	1	1	1	0	0	0	0	0
0	1	1	0	0	0	1	1	1	0	1	0	1	0
0	0	1	1	0	0	1	1	1	0	0	0	1	0
0	0	1	0	0	0	1	1	1	0	1	0	1	0
0	1	1	0	0	0	1	0	1	0	1	0	1	0
0	0	1	0	0	0	1	0	1	0	1	0	1	0
0	0	1	0	0	0	1	1	1	0	1	0	1	1
0	1	1	0	0	0	1	1	1	0	0	0	1	0
0	1	1	0	0	0	1	0	0	0	0	0	1	1
0	1	1	0	0	0	1	1	0	0	1	1	1	0
0	1	1	1	0	0	1	1	1	1	0	0	0	1
0	1	1	1	0	0	1	1	1	0	0	0	1	0
0	0	1	1	0	0	1	1	1	1	1	0	1	0
0	1	1	0	0	0	1	0	1	0	1	0	1	1
1	1	0	1	1	1	0	0	0	1	0	1	0	1
0	1	1	0	0	1	1	0	1	0	1	0	1	0
0	1	1	0	0	0	1	1	1	0	1	0	1	0
0	1	1	0	0	0	1	0	1	0	1	0	1	0
0	0	1	0	0	0	1	1	1	0	1	0	0	0

Table 6: Raw data on 19 tasks for 31 subjects in PCC who were each payed a flat fee for participating in the experiment. The first line contains a list of the 19 priors for the 19 tasks (2 refers to the prior of 1/3 in favor of cage A, 3 refers to the prior of 1/2 in favor of cage A, and 4 refers to the prior of 2/3 in favor of cage A). The second line contains a list of the 19 outcomes (number of N's observed, between 0 and 6). The last 31 lines contain the 19 choices of each of the subjects (choosing cage A is coded as 1, and choosing cage B is coded as 0), for each of the tasks.

2	2	4	4	2	2	4	4	3	3	3	3	2	2	4	4	3	3	3
4	3	2	5	4	3	2	4	6	4	2	2	5	2	3	2	3	2	2
1	0	1	1	0	1	1	0	0	0	1	1	0	0	1	1	1	1	1
0	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	1	0
1	0	1	1	0	0	0	1	1	0	1	1	0	1	0	0	1	0	1
0	0	0	1	0	0	0	1	1	1	0	0	1	0	1	0	0	0	0
1	0	1	1	0	0	1	1	1	1	0	0	1	0	0	0	1	0	0
1	0	0	1	0	0	1	1	1	1	0	0	0	0	1	0	0	0	0
1	0	0	1	1	0	0	1	1	1	0	0	1	0	1	0	0	0	0
1	0	0	1	1	0	0	1	1	1	0	0	1	0	0	0	0	0	0
1	0	0	1	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0
0	0	0	1	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0
0	0	1	1	0	0	1	1	1	1	1	0	0	1	0	0	1	1	1
1	0	0	1	0	0	0	1	1	1	0	0	1	0	1	0	0	0	0
0	0	0	1	0	0	0	1	1	1	0	0	1	0	1	1	0	0	0
0	1	1	1	0	0	0	1	1	1	1	0	1	0	0	0	0	0	0
1	0	0	1	0	0	0	1	1	1	0	0	1	0	1	0	0	0	0
1	0	0	1	0	0	0	0	1	1	0	0	1	0	0	0	0	0	1
0	0	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	0	1
1	0	0	1	0	0	0	1	1	1	0	0	1	0	1	0	0	0	0
0	0	0	1	0	0	0	1	1	1	0	0	0	0	1	0	0	0	0
1	0	0	1	1	0	0	0	1	1	1	1	1	0	0	0	0	0	0
1	0	0	1	0	1	0	1	1	1	0	0	1	0	1	0	1	0	0
0	0	0	1	1	1	0	1	0	0	0	0	0	0	1	0	0	0	1
1	0	0	1	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0
0	0	0	1	1	0	0	0	1	0	0	1	0	0	1	1	0	1	0
1	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	0	0
0	1	0	1	1	0	1	1	1	1	0	1	1	0	0	0	0	0	0
1	0	0	1	1	0	0	1	1	0	0	0	1	0	1	0	0	0	0
0	0	1	0	0	0	1	1	0	0	1	1	0	1	0	1	0	0	0
0	0	0	1	1	0	0	1	1	1	0	0	1	0	1	1	0	0	0
1	1	1	1	0	0	0	1	1	1	0	0	1	1	0	0	0	0	1
1	0	0	1	1	0	0	1	1	1	0	0	1	0	0	0	0	0	0

Table 7: Raw data on 19 tasks for 17 subjects in CSULA who were each payed a bonus of $10 if they made the correct choice on a randomly drawn task. The first line contains a list of the 19 priors for the 19 tasks (2 refers to the prior of 1/3 in favor of cage A, 3 refers to the prior of 1/2 in favor of cage A, and 4 refers to the prior of 2/3 in favor of cage A). The second line contains a list of the 19 outcomes (number of N's observed, between 0 and 6). The last 17 lines contain the 19 choices of each of the subjects (choosing cage A is coded as 1, and choosing cage B is coded as 0), for each of the tasks.

2	2	4	4	2	2	4	4	3	3	4	4	4	3	4	4	4	2	3
5	3	5	6	5	4	2	4	5	3	6	4	4	3	1	4	3	4	4
0	0	0	1	1	1	0	1	1	0	1	1	1	0	0	1	0	1	1
1	0	1	1	1	0	0	1	1	0	1	1	1	0	0	1	0	0	1
1	1	1	0	1	0	1	1	0	0	1	0	0	1	0	1	0	0	1
1	0	1	1	1	0	0	1	1	0	1	0	0	0	0	1	0	0	0
0	0	1	1	0	0	0	1	0	0	1	1	0	0	0	1	0	0	1
0	0	1	1	0	0	0	1	1	0	1	1	1	0	0	1	1	0	1
1	0	1	1	1	1	0	1	1	0	1	1	1	0	0	1	0	1	0
1	0	1	1	1	0	0	1	1	0	1	1	0	0	1	1	1	0	1
1	0	1	1	1	0	0	1	1	0	1	1	1	0	0	1	0	0	1
0	0	1	1	0	0	0	1	1	0	1	1	1	0	0	1	1	0	0
1	0	1	1	1	0	0	1	1	0	1	1	1	0	0	1	0	0	1
1	0	1	1	1	1	0	1	1	0	1	1	0	0	0	1	0	1	1
1	0	1	1	1	0	0	1	1	0	1	1	1	1	0	1	1	0	1
0	1	0	0	1	0	1	1	1	0	1	0	1	1	1	0	0	0	1
1	0	1	1	1	0	0	1	1	0	1	1	1	0	0	1	1	0	1
1	0	1	1	0	0	0	1	1	1	1	0	1	1	0	0	1	0	1
1	0	1	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	1

Table 8: Raw data on 19 tasks for 20 subjects in PCC who were each payed a flat fee for participating in the experiment. The first line contains a list of the 19 priors for the 19 tasks (2 refers to the prior of 1/3 in favor of cage A, 3 refers to the prior of 1/2 in favor of cage A, and 4 refers to the prior of 2/3 in favor of cage A). The second line contains a list of the 19 outcomes (number of N's observed, between 0 and 6). The last 20 lines contain the 19 choices of each of the subjects (choosing cage A is coded as 1, and choosing cage B is coded as 0), for each of the tasks.

2	2	4	4	2	2	4	4	3	3	3	3	2	2	4	4	3	3	2
3	1	3	3	4	4	4	6	5	5	4	2	4	4	1	2	6	4	2
0	1	0	1	1	0	0	1	1	0	0	0	0	1	0	0	0	1	0
1	0	1	0	0	0	1	1	1	0	0	0	0	1	0	0	1	1	0
0	0	1	0	1	0	1	1	0	1	1	0	1	0	1	1	1	1	0
0	1	0	0	1	1	1	1	1	0	0	1	0	0	0	1	0	1	0
0	1	1	0	0	1	0	1	1	1	1	0	1	1	0	0	1	0	0
0	0	1	0	0	1	1	1	1	1	1	0	0	0	0	0	1	1	0
0	1	1	0	0	0	0	0	1	1	1	0	1	0	1	1	1	0	1
0	0	1	1	0	0	1	1	1	1	1	0	0	0	0	1	1	1	0
0	0	1	0	1	0	1	1	1	1	1	0	0	0	1	0	1	0	0
0	0	1	1	1	0	1	1	1	1	1	0	0	1	0	0	1	1	0
0	0	0	0	1	1	1	1	1	1	0	0	1	1	0	0	1	1	0
0	0	1	0	1	1	1	1	1	0	0	0	0	1	0	0	0	1	0
0	0	0	0	1	1	1	0	0	1	1	0	1	1	0	1	0	0	1
0	0	0	1	1	0	1	1	1	1	0	0	1	1	0	1	0	1	0
1	1	0	1	1	0	1	1	1	1	1	0	1	1	0	1	0	1	0
0	0	0	1	1	0	1	1	1	1	1	0	1	0	0	0	1	1	0
1	0	1	1	1	1	1	1	1	1	0	0	0	0	1	0	1	0	0
0	0	1	1	0	0	1	1	1	1	1	0	0	0	1	1	1	1	0
0	0	0	1	0	0	1	1	1	1	1	0	0	0	0	0	1	1	0
0	0	0	0	1	1	1	1	1	1	1	0	1	1	0	0	1	1	0

10 Appendix C

Instructions

Name: ____

Social Security Number: ______________________________________

The experimenters are trying to determine how people make decisions. We have designed a simple choice experiment and we shall ask you to make decisions at various times. The amount of money (DESIGN I: will not) (DESIGN II: will) depend on how good your decisions are. (DESIGN I: At the end of the experiment you will receive $7) (DESIGN II: At the end of the experiment, we shall randomly choose one of your decisions and if it was correct, you will receive $15, and if it was incorrect, you will receive $5).[4]

If you look at the person in the front of the room, you will see that he or she has three randomizing devices, otherwise known as bingo cages, which are designated as cage A, cage B, and cage X. Inside both cage A and cage B are six balls, some of which are marked with an N and some with a G. Cage A has four N's and two G's and cage B has three N's and three G's. Inside cage X there are six balls numbered one, two, three, four, five, and six.

The experiment will proceed as follows. First, we shall ask you to select one individual as a monitor to watch the procedures, to examine the equipment, and to make sure that the experimenters really are doing what they say they are doing. The monitor should check the truthfulness of what the experimenter says, but other than that may not communicate any information to you in any way. If the monitor communicates any other information, he or she will be asked to leave without payment. The monitor will receive (DESIGN I: $7) (DESIGN II: $10) for his or her efforts.

(PICK VOLUNTEER)

Now cages A and B are put behind this screen and we spin cage X. Before each run we will tell you that if certain numbers come up, we will choose cage A, and other wise we will choose cage B. For example, if 1, 2, 3, or 4 are drawn, we will pick cage A; if a 5 or 6 is drawn, we will pick cage B. After drawing from cage X, the monitor will be asked to choose the appropriate cage (A or B) from behind the screen. Now the experimenter will make six draws from the cage, replacing the drawn ball each time. He will write the results of the draws on the board. You will be asked to indicate on your answer sheet whether you think the draws came from cage A or cage B.

[4]Legend: DESIGN I = Subjects are payed a flat fee. DESIGN II = Subjects are payed a bonus if a randomly chosen choice was correct.

Answer Sheet

Cage A has 4 N's and 2 G's.
Cage B has 3 N's and 3 G's.

Run No. 0 In this run there are —— chances out of 6 for choosing cage A and —— chances for choosing cage B.

Record the results for this run: — — — — — —.

Circle the one you think the balls came from. Cage A Cage B

Run No. 1 In this run there are —— chances out of 6 for choosing cage A and —— chances for choosing cage B.

Record the results for this run: — — — — — —.

Circle the one you think the balls came from. Cage A Cage B

Run No. 2 In this run there are —— chances out of 6 for choosing cage A and —— chances for choosing cage B.

Record the results for this run: — — — — — —.

Circle the one you think the balls came from. Cage A Cage B

...

...

...

References

Akaike, H. 1974. A new look at the statistical idenitification model. *IEEE Transactions on Automatic Control* 19:716–723.

Allais, M. 1953. Le comportement de l'homme rationel devant le risque,critique des postulates et axiomes de l'ecole americaine. *Econometrica* 21:503–546.

Bell, D. 1982. Regret in decision making under uncertainty. *Operations Research* 30:961–981.

Birnbaum, M., G. Coffey, B. Mellers, and R. Weiss. 1992. Utility measurement:configural-weight theory and the judge's point of view. *Journal of Experimental Psychology:Human Perception and Performance* 18(2):331–346.

Bostic, R., R. Herrnstein, and D. Luce. 1990. The effect on the preference-reversal phenomenon of using choice indifference. *Journal of Economic Behavior and Organization* 13:193–212.

Boylan, R. and M. El-Gamal. 1993. Fictitious play: A statistical study of multiple economic experiments. *Games and Economic Behavior* 5(2):205–222.

Cheeseman, P. 1988. Autoclass II conceptual clustering system. *Proceedings of Machine Learning Conference* pages 54–64.

Chew, S. 1983. A generalization of the quasilinear mean with applications to the measurement of income inequality and decision theory resolving the allais paradox. *Econometrica* 51:1065–1092.

Edwards, W. 1982. Conservatism in human information processing. In Kahneman, D., P. Slovic, and A. Tversky, eds., *Judgment Under Uncertainty: Heuristic and Biases*, pages 359–369. Cambridge University Press, Cambridge.

El-Gamal, M., R. McKelvey, and T. Palfrey. 1993. A Bayesian sequential experimental study of learning in games. *Journal of the American Statistical Association* 88:428–435.

Ellsberg, D. 1961. Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics* 75:643–669.

Gigerenzer, G., H. Ulrich, and H. Kleinbolting. 1991. Probabilistic mental models: A brunswikian theory of confidence. *psychological Review* 98(4):506–528.

Goldstein, W. and H. Einhorn. 1987. Expression theory and the preference reversal phenomena. *Psychological Review* 94(2):236–254.

Grether, D. 1980. Bayes rule as a descriptive model: The representativeness heuristic. *Quarterly Journal of Economics* pages 537–557.

Grether, D. and C. Plott. 1979. Economic theory of choice and the preference reversal phenomenon. *American Economic Review* 69:623–638.

Kahneman, D. and A. Tversky. 1972. Subjective probability: A judgment of representativeness. *Cognitive Psychology* 3:430–454.

Kahneman, D. and A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47:263–292.

Lichtenstein, S. and P. Slovic. 1971. Reversals of preferences between bids and choices in gambling decisions. *Journal of Experimental Psychology* 89:46–55.

Loomes, G., C. Starmer, and R. Sugden. 1989. Preference reversal:information-processing effect or rational non-transitive choice? *Economic Journal* 99:140–151.

Loomes, G. and R. Sugden. 1987. Some implications of a more general form of regret theory. *Journal of Economic Theory* 41:270–287.

Machina, M. 1982. Expected utlity analysis wothout the independence axiom. *Econometrica* 50:277–323.

Mellers, B., L. Ordonez, and M. Birnbaum. 1992. A change-of-process theory for contextual effects and preference reversals in risky decision making. *Organizational Behavior and Human Decision Processes* 52:331–369.

Payne, J. 1982. Contingent decision behavior. *Psychological Bulletin* 92(2):382–402.

Quigen, J. 1982. A theory of anticipated utility. *Journal of Economic Behavior and Organization* 3:323–343.

Rissanen, J. 1978. Modeling by shortest data description. *Automatica* 14:465–471.

Rissanen, J. 1987. Stochastic complexity. *Journal of the Royal Statistical Society B* 49(3):223–239 and 252–265.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6,2:461–464.

Tversky, A. and D. Kahneman. 1972. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* 51:207–232.

Tversky, A. and D. Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185:1124–1131.

Tversky, A., S. Sattath, and P. Slovic. 1988. Contingent weighting in judgement and choice. *Psychological Review* 95(3):371–384.

Wallace, C. and D. Boulton. 1968. An information measure for classification. *computer Journal* 11:185–195.

Wallace, C. and P. Freeman. 1987. Estimation and inference by compact coding. *Journal of the Royal Statistical Society B* 49(3):240–265.

Yaari, M. 1987. The dual theory of choice under risk. *Econometrica* 55(1):95–116.